We would like to thank the handling editor for taking the time to review the manuscript and the suggestions. Please see our responses and updates to the text for each comment below.

Based on my assessment the paper is suitable for GMD, but it still has a few shortcomings that I'd like to see addressed:

- the paper is very descriptive at times (list of numbers over several sentences). Please try to extract and emphasize the relevant information instead.

We revisited the result sections and focused on the main points to improve readability. The values were extracted into tables and moved to the supplement material. Please see the updates to the results in section 5.2 and the responses on L 275 and L 290 below.

- a lot of emphasis is given to the advantage of energy balance methods (iSnobal) over older temperature index methods (Snow-17). Yet nowhere is the performance of both models compared to each other with observations as reference. Section 5.5 only provides a qualitative comparison. Why? I think it would give more substance to the paper to actually show that iSnobal has other qualities than "being applicable without calibration".

The different nature of Snow-17 and HRRR-iSnobal makes it hard to use observation metrics that are fair in a 1-to-1 comparison. The two models use different input data (calibrated precipitation versus numerical weather prediction modeled precipitation), are different in spatial extent and representation (elevation-based hydrological response units (HRU) vs. user-defined gridded spatial resolution), and use different principles (temperature index-based vs. physics-based) to arrive at snow water equivalent (SWE) outputs. SWE is the most relevant variable from an operational hydrology standpoint and hence the focus metric in section 5.5. Adding additional observations, that would be considered 'truth' between the two models, are currently not available. Snow depths, for instance, can not be used with Snow-17 as it is an aerial lump sum model. Available SWE products (i.e. by ASO) use modeled snow density and can't be considered as true either. With this difference and scarcity of reference values, we would like to make the case that section 5.5 is a reasonable approach to comparing the two. Here, we used the HRU of Snow-17 to categorize the SWE outputs between the two models to overcome some of the different model principles.

To highlight the advantages of a physically based model over a temperature index model, we revisited the discussion section 6.3 that present the arguments for application beyond the voided need of calibration:

In summary, physically based models remove the dependence on long-term historical calibration data, reduce the need for user intervention due to parametrization issues, require less model domain specific user knowledge, and better represent physiographic influences. All these factors combined result in improved scalability across different seasonal and terrain-dependent snowpack dynamics. Gradually adding these models into operational settings, with architectures presented in this work, can enhance snowpack information in response to current environmental perturbations and expand the ability to adapt to current and future water supply forecast needs.

- to this latter point: how was iSnobal calibrated?

iSnobal does not use any data for calibration and is controlled via a central configuration file. The configuration gives the user options to customize parameters that might be different for a model domain. The parameters include options such as soil temperature, height of wind, or maximum grain size. For this run, we kept all values to the default values. We added a new section to document this:

3.2 Model Configuration

Each component of the model (Katana, SMRF, AWSM) is controlled by a configuration file that allows the user to customize parameters used during execution. For instance, when distributing forcing data from HRRR with SMRF, the user can change values for minimum and maximum air temperature, maximum and minimum snow grain sizes when calculating albedo decay, or the height above ground when calculating turbulent fluxes from the wind data. An overview of the possible options and default values for each component is available in the respective published documentation. For this application, no parameters were changed from the default values and the configuration files are available on the GitHub (https://github.com/UofU-Cryosphere/isnoda).

Inline Comments

L 49 - This is really counter intuitive. The more data the better in general. I think I know what you want to say, but it should be formulated differently.

Please also refer to litterature when making statements such as "temperature index models are less valid in a warming world". I'm actually unconvinced this is really that obvious, even if ofc it makes sense to move towards more physically based approaches.

We revisited this paragraph and removed the sentence in this section. The core idea of this statement is now combined with the paragraph starting on line 81 and also addresses the comment on L 121. The updated paragraph with the sentence removed:

Presently, a subset of the hydrologic forecast agencies in the United States use temperature-index models, such as SNOW-17 (Anderson 1976), which have historically performed well in operational settings while requiring few meteorological observations (Franz et al., 2008). In principle, SNOW-17 calculates the snowmelt using the correlation between air temperature and available net solar radiation melt energy and a calibration factor, which increases as the melt period progresses (Anderson, 1976; Franz et al., 2010). The best model predictions are with domain-specific calibration parameters from historical data with the modeled year following the snow accumulation and melt conditions from the past (He et al., 2011). Once conditions depart from the historical average, such as lower snow albedo from highly variable inter-annual dust deposition events (Bryant et al., 2013), the SNOW-17 model forecast errors increase and require significant forecaster interaction to account for the variable conditions. One effort to improve the accuracy of SNOW-17 applied the Bayesian Model Averaging method across an ensemble of twelve snow models, each consisting of different components from SNOW-17 (Franz et al., 2010). Although the results improved compared to running SNOW-17 as a standalone application, the setup was only tested at the 1d point scale *and required different weights for the individual models between test locations. The increased complexity makes the method challenging to apply across larger spatial scales in daily operations.*

L 81 - available

Fixed

L 94 - Again: reference?

We addressed the comment on L 81 together with this paragraph. The revised statement is now:

Freshwater supply forecasting in this region is done by the Colorado Basin River Forecast Center (CBRFC), part of the National Weather Service (NWS) in the United States of America. The CBRFC uses SNOW-17 as part of their operational water availability forecasting model and faces increased challenges with the near-term observed and predicted seasonal snow changes. Using the historic model calibration records (30 to 40 years) to derive SNOW-17 parameters does not fully reflect the climate variability in the recent decade and the long-term data will continue to be less representative in the future (Musselman et al., 2017). Incorporating the different timing and magnitudes of snowmelt requires updated methods.

L 121 - Is this accronym really necessary?

We removed the acronym throughout the text.

L 215 - What about model calibration??? This needs to be explained.

Please see our response in the main comment regards model configuration.

L 275 - These are really a lot of numbers and is nearly not readable. Synthesis?

We revisited the entire paragraph, extracting the relevant information, and moved most of the numbers into the Supplement as Table 1. The revisited paragraph:

At the Butte site, the early survey flights across all years had HRRR-iSnobal consistently higher than ASO. However, the measured values at this SNOTEL site agreed more with the 2x2 spatial grid snow depths of HRRR-iSnobal (difference ranging from -0.19 m to +0.13 m for all years) versus ASO (-0.48 m to -0.07 m). The 2018 late survey flight values agreed across all three sources for Butte, where snow was completely melted. In 2019, HRRR-iSnobal still had snow present, with ASO and SNOTEL showing complete melt out. Schofield Pass had good agreement across all snow depth values for the early 2018 flight (HRRR-iSnobal difference +/- 0.04 m; ASO +0.05 m and -0.16 m), whereas the above-average 2019 season had lower values in HRRR-iSnobal with ASO capturing the SNOTEL depth value. The 2018 late survey flight had iSnobal-HRRR and ASO above the Schofield Pass value, with HRRR-iSnobal and ASO having a similar spatial range (HRRR-iSnobal: 0.34 m; ASO: 0.25 m). In 2019, the late survey flight captured the site value (1.50 m) in HRRR-iSnobal (range from 0.84 m to 1.61 m) and ASO (1.25 m to 1.68 m). The Upper Taylor site location was only included in the early 2019 flight and confirmed the strong overestimation by HRRR-iSnobal (1.61 m to 2.06 m) as the ASO snow depths values (1.40 m to 1.61 m) agreed with the SNOTEL depth (1.45 m). The agreement between ASO and SNOTEL and overestimation by HRRR-iSnobal was in both late survey flights for 2018 and 2019 at the Upper Taylor location, as the snow depths approached 0 m. An overview of snow depth values across all ASO flight years is given in Supplement Table ST2. Overall, using ASO as an additional snow depth reference data set at discrete point locations in the model domain showed no consistent over or under-simulation for HRRR-iSnobal across the years.

L 287 - How about differences between ASO and SNOTEL? How can these be explained?

We would like to keep the paper focus on the HRRR-iSnobal performance and not add an evaluation of ASO data. Assessing this difference is indeed interesting and would warrant a paper by itself.

L 290 - Again - is there a way to synthesise these results in a bigger picture which is more than a list of numbers?

The paragraph was revisited and numbers reduced as ranges across the HRUs. This hopefully helps the reader to focus on the summary given. The revisited sentence:

Within lower, middle, and upper HRU elevations, the widest range in the Δ snow depth distribution was consistently found at the upper HRU (Standard Deviation (SD) between 0.53 m and 0.99 m across the flights), while the the lower HRU (SD 0.03 m to 0.55 m) and middle HRU showed similar ranges (SD 0.22 to 0.53 m).

L 304 - Probably because there is no snow left?

We updated the sentence to:

The closest agreement between modeled and ASO snow depths was on the east to south-facing slopes during late flights, as most snow was melted at that time of the season.

L 314 - Could also be due to groundwater travel time in the basin?

In a snow dominated basin like the East River, the rising and receding limb of the hydrograph is strongly dominated by snowmelt, although certainly there is groundwater contribution as well (Carroll et al., 2018). To clarify the point we are trying to make with this comparison (that the receding limb of simulated SWI should precede, not correspond with, the receding limb of the hydrograph) we updated this result section to:

The watershed 7-day moving average of simulated HRRR-iSnobal SWI followed the hydrograph timing and magnitude pattern measured at the stream gauge (Figure 8). During the annual snowmelt pulse, the HRRR-iSnobal SWI stayed higher relative to the measured surface water at the gauge. This behavior was expected, as a portion of SWI is taken up by the ground, plants, and atmosphere before reaching the gauge. The later than observed simulated HRRR-iSnobal snow disappearance dates were also apparent in this comparison. The receding limb of simulated SWI should come before the receding limb of observed discharge, rather than at the same time as is seen in this comparison, because it takes time for the SWI to flow down to the gauge. Still, the general patterns and magnitude are promising, as there is no under-forecasted SWI at any point in the melt season.

Figure 5 - between HRRR-iSnboal and ASO. Make clear that you are using the MODEL - REF convention

The caption was updated to the REF convention.

Figure 10 - Move this to supplement? not very informative figure...

We moved this figure to the supplement as Figure S5.